IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 20, 2020, accepted April 27, 2020, date of publication May 4, 2020, date of current version May 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992064

A Combined Approach for Customer Profiling in Video on Demand Services Using Clustering and Association Rule Mining

SINEM GUNEY[1], SERHAT PEKER[2], AND CIGDEM TURHAN[1]
[1]Department of Software Engineering, Atilim University, 06830 Ankara, Turkey
[2]Department of Management Information Systems, İzmir Bakirçay University, 35665 İzmir, Turkey

Corresponding author: Sinem Guney (sinemmguney@gmail.com)

ABSTRACT The purpose of this paper is to propose a combined data mining approach for analyzing and profiling customers in video on demand (VoD) services. The proposed approach integrates clustering and association rule mining. For customer segmentation, the LRFMP model is employed alongside the k-means and Apriori algorithms to generate association rules between the identified customer groups and content genres. The applicability of the proposed approach is demonstrated on real-world data obtained from an Internet protocol television (IPTV) operator. In this way, four main customer groups are identified: ''high consuming-valuable subscribers'','' less consuming subscribers'','' less consuming-loyal subscribers'' and ''disloyal subscribers''. In detail, for each group of customers, a different marketing strategy or action is proposed, mainly campaigns, special-day promotions, discounted materials, offering favorite content, etc. Further, genres preferred by these customer segments are extracted using the Apriori algorithm. The results obtained from this case study also show that the proposed approach provides an efficient tool to form different customer segments with specific content rental characteristics, and to generate useful association rules for these distinct groups. The proposed combined approach in this research would be beneficial for IPTV service providers to implement effective CRM and customer-based marketing strategies.

INDEX TERMS Customer segmentation, data mining, clustering, association rules, RFM model, VoD services.

I. INTRODUCTION

Recently, increasing the quality of any given network has become the priority of all broadcasting firms. Among these, one can refer to the VoD – video on demand, otherwise known as Pay-TV - business model, which offers broadcast transmissions of media content to subscribers paying for the services [1]. The Internet Protocol Television (IPTV) supports VoD, concentrating more on content and system quality in accordance to the subscribers' choices [2]. With added expansion on the Web, VoD and live TV services of this nature are becoming even more popular. VoD, as a business model for IPTV, has promoted the consumption of TV and video content by subscribers, who can select and rent – oftentimes, free of charge in certain promotional periods - among pre-recorded material, even pause content, rewind and play it back [3]. Additionally, the services provided allow for access and storage of content using a set-top box (STB), laptops, mobile phones or web interfaces with real-time downloading and streaming. In this way, VoD allows for both wide-ranging and quality services to be rendered to customers. For now, improving the VoD market also entails the use of smart devices by end-users to access television as IPTV. According to studies, the VoD market is expected to increase from USD 38.9 billion in 2019 to USD 87.1 billion by 2024, progressing at an average rate of 17.5% [4]. Other reports state that the number of VoD subscribers worldwide now exceeds one billion, and that this figure is predicted to reach 1.1 billion by 2024. In this sense, these services are clearly shown to remain appealing to audiences in the upcoming years [5].

With the IPTV's basic role as one service provider with several subscribers, it has become more important for the firms to determine their subscribers' values and demands. Parallel to this issue, CRM (Customer Relationship Management) involves methodology and practices for making use of customer data in business, where decisions are made based on such data to gain competitive advantage over the rest of the sector [6]. Obviously, effective use of IT tools can increase the level of success in CRM processes, and applying algorithms and statistical methods to the customer

The associate editor coordinating the review of this manuscript and approving it for publication was Noor Zaman.

data for more personalized marketing becomes, as such, evident for analytical CRM [7].

Customer analysis is carried out using data mining techniques to determine CRM strategies and increase customer value. To elaborate on these concepts, clustering is employed as a data mining technique commonly applied for customer segmentation [8]. The latter is an important CRM activity to assist service providers in gaining a better understanding of their subscribers' needs and behaviors by introducing improved products and services [9]. Such insight is a key factor in gaining a competitive advantage through increasing customer satisfaction and meeting their needs [10]. This is even more so as different customer profiles constitute a far greater source of valuable data for CRM applications [11]. As part of the process, these categories call for proper segmentation so as to devise cheaper and, yet, profit-generating strategies in the long run [12].

Among different data mining techniques, clustering methods are particularly helpful for analytical CRM tasks, such as customer segmentation and profiling based on similar attributes and shared values among users. Such clustering makes use of the RFM model as well [13]. Short for recency, frequency and monetary, this model delves into customer behavior and individual's characteristics [14]. In data mining technologies, association rule mining and clustering algorithms are also employed to analyze various enterprises for business analytics. In this discipline, the Apriori algorithm can help uncover the relations among different sets throughout a series of exchanges. Data mining, embedded within other CRM tools, is used for analyzing customers - with the ultimate goal being to increase their satisfaction through better marketing decisions, retaining the client base, identifying specific groups, and adjusting the decisions to their needs. Nonetheless, it is fair to state that, by far, the research conducted solely on VoD service providers and their subscribers could lack in-depth analyses and a more through and all-inclusive approach to user classification and, accordingly, strategizing.

According to above mentioned evidence, it is very important for service providers to keep their subscribers for a longer period of time – which is the key to reduce costs and increase their satisfaction based on a better understanding of their needs and expectations. Customers who have common features with IPTV customer segmentation are divided into specific groups, and IPTV customer profiling also serves to depict customer behavior. Since certain behaviors are known, appropriate services are developed for the customers and a better service is provided to the customers. Data mining techniques are also necessary to use appropriate methods with a good concept understanding for both IPTV customer segmentation and profiling. In interpreting customer behaviors from different data sources, data mining technologies allow IPTV customers to be profiled, and IPTV service providers can get to know customers better and offer better services, thereby increasing customer satisfaction.

Despite the fact that customer profiling has repeatedly been observed in other application domains, to the best of our knowledge, no studies have attempted to profile subscribers in VoD service providers. Therefore, this present study is endeavoring to fill this gap and aims to develop a customer profiling scheme by using the real-life VoD service data of IPTV subscribers. In detail, it proposes a combined approach for customer profiling in VoD services using clustering and association rule mining techniques. For this purpose, the LRFMP model is employed to determine the customer values, to which the k-means clustering algorithm is subsequently applied for customer profiling. Then, customer VoD rental preferences are extracted to elicit the relationship between content types and different customer profiles based on association rule mining.

This research makes the following contributions. First, it proposes a novel hybrid approach of combining clustering and association rule mining techniques for profiling subscribers in VoD services. Second, serves a valuable reference for researchers, who are willing to examine customer behaviors in VoD services. Finally, the present study provides VoD service providers with important implications for different customer types in implementing effective marketing and CRM strategies.

This paper is organized as follows: Section 2 reviews the relevant literature, Section 3 presents the combined approach. Section 4 provides the application of the proposed approach. Finally, conclusions are presented with limitations and suggestions for future works in Section 5.

II. LITERATURE REVIEW

A. CUSTOMER RELATIONSHIP MANAGEMENT (CRM)

The literature points to CRM decision tools to employ data mining to customer databases [15]. Intended for customer-based systems, it analyzes users and helps investigate such data so as to predict their possible behavior and attract new customers [16]. Research points to the CRM process as the management of potential, existing and old customer interactions [17]. On this basis, those within certain groups have different expectations and, whereas it is difficult to manage a large number of customers, users sharing similar behavior are grouped together. To this end, service providers identify the variables necessary to group the customers by means of the right variables [18].

The competition in the digital market is a customer-dependent subject, for which new technologies are used to clearly categorize customers [19]. Once these groups are formed, one can obtain the necessary recommendation rules for CRM purposes [20]. Later, the CRM is combined with the LRFMP model to classify users.

B. CUSTOMER SEGMENTATION

As a matter of course, the unavoidable diversity in customer preferences and behaviors poses many difficulties for the sector to individually and adequately respond to each.

For this reason, customer segmentation separates users based on similarities and according to marketing goals or other criteria [21]. These days, this subject is gaining even more attention focused on its significance for competitiveness.

As far as segmentation is concerned, it allows firms to identify customer groups and determine their needs and expectations, respond to those demands, and increase revenue while attracting new and potential users with relevant marketing strategies [22], [23]. In this way, the literature also mentions customer segmentation as a contributing factor for different marketing strategies. Specifically, researchers divide customers' variables into two groups: general and product-oriented [13].

In this study, the product-specific variables are mainly employed to help identify IPTV subscribers' rental and consumption habits. In this vein, RFM has been known as the most effective model in determining the purchasing habit of customers, their behavior analysis and ultimate segmentation [24].

C. RFM MODEL AND ITS VARIATIONS

The model was initially proposed by Huges in 1994 to measure the values related to customer purchasing behavior [25]. RFM stands for the following variables: Recency, to provide information on the time elapsed since the last purchase of the customer; Frequency, as to the total purchase and customer loyalty within a given period; and Monetary, to refer to the average amount spent over a certain period of time [26]. On top of this, the application of the model with data mining techniques allows for maintaining the already-existing customers, increasing customer loyalty, determining needs and gaining new customers [27]. Briefly put, it helps in accurately identifying customer behavior, providing special service to different groups, and reinforcing the relations between service providers and users [12].

Later studies also determined the length of customer involvement with the addition of the length (L) parameter [28]–[30], [31]. Identifying customer loyalty in this way and upon the length parameter has become more important today [32]. Other such parameters include the group (G) parameter to address product category information as well [33]. As to the present study, customer value is determined with the LRFMP model – as the abbreviation implies – based on length and periodicity parameters added to the RFM model previously employed in retailing [34]. This proposed LRFMP model is intended to better identify the characteristics and regularity of customers having different purchasing habits.

D. CLUSTERING

This technique separates individual data objects based on their differences, and it isolates them in groups based on their characteristics [35]. The literature offers numerous such techniques as per the requirements to define cluster variables and to separate clusters with identical classifications [36]. Among many clustering techniques, the k-means is the most popular algorithm used in clustering with its fast and easy implementation. In particular, it is common in various research fields such as data mining, statistical data analysis and customer segmentation [37], [38]. However, to do so, the k-value needs to be specified before the k-means algorithm is applied. After this specification, k center points are chosen arbitrarily in the algorithm, and the data is assigned to a cluster based on the closest center point and by measuring the distance between each data and the randomly chosen centers. Then, another center point is chosen for each cluster, and clustering is carried out based on the new center points. This is repeated to the point that all centers become stable [39]. Consequently, the k-means algorithm improves the variance between clusters and decreases inter-group variance to classify users [40]. When investigating the users' behavior, segmentation also calls for the determination of the actual purchase habits [41]. In doing so, the relationship among the number of segments, the number of viewers, and the number of optimum clusters are applied to identify the precise preferences of subscribers as well as their overall characteristics and attributes [42]. In a sense, segmentation entails the grouping of users sharing similar attributes to support decision-making for better marketing strategies and achieving the future targets set by service providers [43].

E. ASSOCIATION RULE MINING

In business and elsewhere, clustering and association rule methods are commonly used to define patterns in decision-making based on the data mining models. Association rule mining was proposed by Agrawal *et al.* in 1993 [44]. The Apriori algorithm generates candidate (k-size) patterns incrementally and recursively to calculate and combine the frequent (k-1 sized) patterns [45]. Although the technique is mostly used in market basket analysis examples, it is also preferred in determining various content types with categorical data [46].

The basis of the Apriori algorithm is an iterative structure, so that frequent element sets can be detected [47]. For the algorithm to be applied, the data set must be both transactional and categorical, while the directions of the variables in the data set have to be identifiable [48]. Apriori reveals the relationships among different components and, hence, it is applied in many domains such as retail shopping, credit card transactions, telecommunication service purchases, banking, insurance, and health care – in the latter case, in the form of patient histories [49]. It has three main measures: support, confidence and lift; where support is an indication of how frequently the data appears in the given dataset, confidence is the measure of how often the rule has been found to be true, and lastly lift is a correlation analysis measure to identify the relation between antecedent and consequent [50].

III. PROPOSED COMBINED APPROACH

This study proposes a combined approach using data mining methods for clustering and rule analysis to examine and analyze subscribers' VoD behaviors in the IPTV sector.

Figure 1 shows an outline of the steps of the proposed approach. As illustrated there, the proposed approach includes three phases, namely preparing the dataset, segmenting the customers, and creating the association rules. These steps are described below:

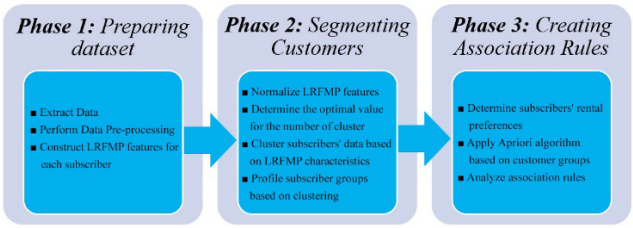

FIGURE 1. The framework of the proposed approach.

Phase 1 starts with extracting data from different sources. Then, data cleaning and pre-processing are performed for the purpose of analysis. In the next step, the LRFMP variables are constructed for each IPTV subscriber. Phase 2 includes the standardization of these variables, clustering and determining the subscriber groups. The most often used scaling technique, min-max normalization, is applied between 0 and 1 to reduce the potential effects of variable differences. Then, it is time to determine the proper number of cluster and customer profiling based on clustering. In phase 3, association rule mining is employed to determine the subscribers' rental preferences. After that, the Apriori algorithm is applied to customer groups to recommend new content types according to the groups. In this way, the proposed combined approach addresses both the IPTV subscribers' profiling by clustering and determining their rental preferences by association rule mining. Finally, the likely rental contents to be favored by these subscribers are predicted based on the customer groups to devise appropriate marketing strategies for adoption.

A. LRFMP MODEL

In the proposed approach, the LRFMP model, as described by Peker *et al.* (2017), is employed to segment IPTV subscribers based on their content rental behaviors. The detailed definitions of these variables adapted to the IPTV domain are presented in the following:

Length (L): the number of days between the subscriber's first and last rentals;

Recency (R): the number of days since the subscriber's last rental;

Frequency (F): the total number of content rentals by the subscriber during the observation period. In the calculation of this variable, each content rental is counted separately for the cases of multiple content rentals in a day;

Monetary (M): the average amount spent per content rental by the subscriber; and

Periodicity (P): the standard deviation of the subscriber's inter-rental times which refers to time in days between two consecutive rentals belonging to different dates. In the calculation of this variable, multiple content rentals for a certain day are considered as single. In other words, only unique dates are considered.

In light of these explanations, Table 1 represents the rental transactions of a given hypothetical IPTV subscriber. The sample subscriber has rented contents nine times, and the rental dates are sorted by date. The date column indicates the content rental dates of the subscriber, and the IRT values are calculated for each rental after the subscriber's first rental. For example, the inter-rental times for the subsequent rents of the subscriber are calculated in days by finding the difference between second and first rental dates. The content fee column indicates the content price for each rental date. The last date of the time period in the example is set as July 31, 2019 to compute the length and recency variables.

TABLE 1. Rental transactions of a sample subscriber.

Date	IRT	Content Fee
May 17, 2019	na	2.50
May 29, 2019	12	1.80
June 4, 2019	6	3.20
		1.50
June 21, 2019	17	2.25
June 24, 2019	3	1.60
		2.75
July 3, 2019	9	2.40
July 24, 2019	21	1.75

Based on the data in Table 1, the length value, calculated as 68, refers to the days between the subscriber's first rental date (May 17, 2019) and the last rental date (July 24, 2019), indicative of the subscriber's degree of loyalty to the IPTV. The recency value, measured to be 7, refers to the number of days between the subscriber's last rental date (July 24, 2019) and the last date of the time period (July 31, 2019), which implies that the subscriber has recently rented VoD content and that (s)he is more likely to rent in the near future. The total number of content rentals represents the frequency value, which is calculated as 9; the higher this value, the more the subscriber's loyalty to the service provider. The monetary value, the average amount spent by a subscriber per rental of content, is determined as 2.19. Given the low VoD content fees, the subscriber's contribution to the IPTV service provider is insignificant. Lastly, the periodicity value, measured at 6.77, refers to the standard deviation of each IRT duration of the subscriber, thus indicating the degree of rental regularity over the stated period. In this respect, low periodicity values point to the subscriber's renting VoD contents more regularly and periodically. This example clarifies the LRFMP model and exemplifies its application to subscribers' VoD behavior in IPTV industry.

B. CUSTOMER SEGMENTATION

In the present work, customer segmentation is carried out with an effective combination of the LRFMP model and clustering. Each LRFMP parameter is regarded as equally significant with thorough standardization in cluster

analysis [51], [52]. The k-means algorithms minimize the total within-cluster sum of square (WSS) to define the clusters. Using the elbow method, we find the proper number of clusters with the total WSS as a function. As a result of verifying the total change in the cluster - which also means that the WSS value will decrease – WSS for different k values is calculated for choosing the optimal k value. After determining the appropriate cluster number, cluster analysis is performed, and customer segments are profiled based on customer values as per the LRFMP model. Then, the content type information preferred by the customers is included in the process to determine the relationship between the rented content types by association rule analysis - described in detail in what follows.

C. ASSOCIATION RULE MINING

In this study, all rented content types are tagged and named in a G (genre) list as:

$$G = \{action, adventure, animation, biography, comedy, crime, documentary, drama, family, fantasy, history, horror, music, musical, mystery, romance, sci-fi, sport, thriller, TV-shows, war, western\}$$

To begin with, the most rented five genres are selected; then, Apriori algorithm is applied to each in order to find out the relationship with other 4 genres. In this algorithm, each association rule is composed of two different item sets denoted as X and Y [53]; X represents the antecedent (left hand side) item with the most rented content type, whereas Y stands for the consequent (right hand side) item set relevant to X [47]. Here, association rule mining allows for the computation of the probability of rental content types. In the same way, the antecedent and the consequent express the degree of uncertainty about the rule. Support simplifies the number of transactions and contains all items in the antecedent and consequent parts of the rule. It is an indication of how frequently the itemset appears in the dataset. Confidence includes all items in the consequent and the support to the number of transactions, including all items in the antecedent. Confidence represent the frequency in which the rule can be true. Lastly, lift denotes the relation between the antecedent and consequent items. Lift is helpful because it takes into account support as well as the dataset. The association rule has the form of $X \rightarrow Y$ and the formulas of support, confidence and lift measurements are given below [44]:

$$Support\,(X \rightarrow Y) = P\,(X \cap Y) \quad (1)$$

$$Confidence\,(X \rightarrow Y) = P(X|Y) \quad (2)$$

$$Lift\,(X \rightarrow Y) = \frac{P(X|Y)}{P(Y)} \quad (3)$$

The customer rental preferences and possible future rental content types are determined by applying the Apriori algorithm, which allows us to obtain the most preferred content types of IPTV subscribers as well as the potential rental content types they would rent in the future. As a result, by analyzing the customer profiles determined with clustering, content type analysis is conducted, and reliable association rules are obtained to examine the dependencies among the different content genres.

IV. APPLICATION OF THE PROPOSED APPROACH

A. THE CASE COMPANY AND DATA PREPARATION

The service provider in point is one of the major digital broadcasting platforms in Turkey. It began its operations in the early 2010s, and now its video streaming services reach the STB (set-top-box)-based data of the subscribers. In this regard, the original data set was extracted from 277,808 subscribers' STB-based data during a two-year period. The VoD subscriber data is obtained in the form of a spreadsheet, which was subsequently transferred to a Microsoft SQL Server database for analysis. In this dataset, subscribers with a length value of only one rental date and those with zero monetary value were excluded from the study alongside users who had only one content rental – the reason being that, otherwise, the periodicity value would not be calculable due to lack of the IRT measure. In the data preprocessing stage, missing subscribers' information and incorrect transaction records were removed. Therefore, in the end, 195,493 subscribers' data related to rental records were analyzed. In the data set, each subscriber's transaction contains their ID, platform information, rental date, content ID, content price, name, and content type. The variables of the LRFMP model were produced for each subscriber. Accordingly, Table 2 presents the descriptive statistics regarding the maximum, minimum, averages and standard deviation results of the LRFMP variables.

TABLE 2. The descriptive statistics based on LRFMP variable.

	Maximum	Minimum	Average	SD
Length	728	1	220.03	200.2
Recency	727	1	105.43	159.9
Frequency	733	2	14.61	21.8
Monetary	39	0.2	2.51	1.5
Periodicity	502.8	0	33.6	50.3

B. CUSTOMER SEGMENTATION RESULTS

In the study, the LRFMP variables are standardized by using min-max normalization between 0 and 1 prior to clustering. Then, the ideal number of clusters is obtained with the LRFMP values using the k-means algorithm. Figure 2 depicts the WSS results of the k-means algorithm for different number of clusters between 2 and 10. On this basis, the curve in the figure indicates that the WSS value declines sharply until k=4, whereas the number of clusters does not decrease considerably with high number of subsequent clustering. Thus, 4 is chosen as the optimum number of clusters.

Table 3 presents the sample size, the average values of the LRFMP variables, and the scores for each cluster. In interpreting the LRFMP score, the up and down symbols previously

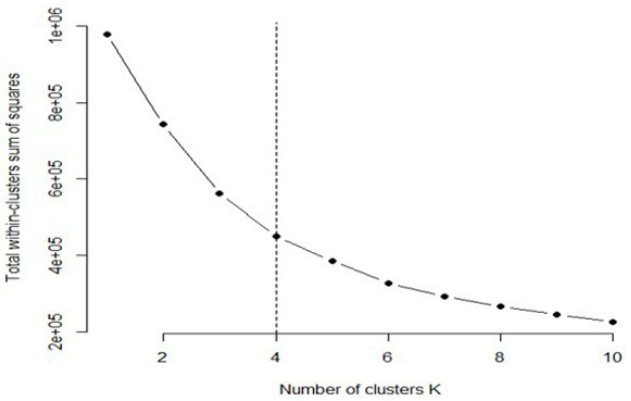

FIGURE 2. WSS values of the k-means clustering method.

TABLE 3. Clustering results for 4 clusters.

Cluster	Sample Size	Avg L	Avg R	Avg F	Avg M	Avg P	LRFMP Scores
Cluster1	18,194	440.46	23.71	66.84	1.51	15.54	L↑R↓F↑M↓P↓
Cluster2	28,761	104.64	445.59	5.96	4.03	17.06	L↓R↑F↓M↑P↓
Cluster3	39,925	482.55	63.14	11.19	2.80	102.41	L↑R↓F↓M↑P↑
Cluster4	108,613	117.17	44.59	9.40	2.17	15.71	L↓R↓F↓M↓P↓
Avg		220.03	105.43	14.61	2.51	33.60	

used in the literature are employed here as well [54]. The technique explains that the L, R, F, M and P values of the cluster are above the aggregate average; hence, the up symbol (↑); otherwise, the down (↓) symbol.

To explain the different customer groups, related profiles are created based on the results obtained in the cluster analysis. Then, customer value [55] and customer relationship matrices are employed to interpret the segmentation results [56]. In all, the study includes 4 clusters with different characteristics as determined in Table 3. Table 4, in turn, represents the names of the subscriber groups, LRFMP scores and the size of each customer group.

TABLE 4. The results of customer groups.

Customer Groups	Name of group	LRFMP Scores	Size of group (%)
1	High consuming-valuable subscribers	L↑R↓F↑M↓P↓	9.30
2	Less consuming subscribers	L↓R↑F↓M↑P↓	14.71
3	Less consuming-loyal subscribers	L↑R↓F↓M↑P↑	20.42
4	Disloyal subscribers	L↓R↓F↓M↓P↓	55.55

Referring to Table 4, one can see that in Cluster1 the length and frequency are greater than the average, and that recency and periodicity are lower than the average. This type of subscribers is, therefore, named ''high consuming-valuable subscribers'' who tend to maintain long-term relationships with the VoD service provider. Hence, as regards marketing strategies and decisions, various promotions and discounts could be provided to this group as it is profitable to the firm. Next, cluster 2 has length, frequency and periodicity values that are lower than the average, and recency which is greater than the average. This type of subscribers are classified as new subscribers and grouped as ''less consuming subscribers''. As a result, IPTV service providers may offer more contents to these subscribers and make the services more attractive to them in the course of time because they tend to rent high-price content. In Cluster 3, the length and monetary values are greater than the average; recency, frequency and periodicity are lower. Henceforth, the group is regarded as ''less consuming-loyal subscribers'' who prefer high-priced contents – which, in turn, can be determined more accurately and increased in terms of its variety to encourage longer subscription with the service provider. The last group, cluster 4, has length, recency, frequency, monetary and periodicity values lower than the aggregate average and, as such, it is named ''disloyal subscribers''. This groups comprises the largest population, whose expectations and characteristics deserve thorough and detailed studying to make them more active.

C. ASSOCIATION RULE MINING BASED ON CUSTOMER PROFILING

In the data set of IPTV subscribers, content types and subcategories of the rented contents were extracted. Each of the subscribers' leased content type transaction is recorded to the database with values of 0 and 1 as shown in Table 5.

TABLE 5. Most rented content type transactions of sample subscriber.

Customer ID	Action	Adventure	Comedy	Drama	Sci-Fi
Customer1	1	1	0	0	0
Customer2	0	0	1	0	0
Customer3	1	0	0	1	1
Customer4	1	1	1	0	0

In the given IPTV subscribers' rental content types, the size of each type is determined with the Apriori algorithm for different subscriber groups. Then, the most popular types or genres are determined to find out about the interrelations among them. The ratio of the number of rentals to the total number of rentals is set as 0.001 as the support value. Additionally, confidence is set to 0.9 to determine the best rules due to the extensive size of the data set. According to these values, the Apriori algorithm is applied to find out the relationship between each customer group and the most rented content types.

For each group, a series of association rules are produced according to the minimum support and the minimum of confidence values. Then, the relations of association rules above these values are identified. In this study, the IPTV subscribers are scanned to generate candidate itemset from the database [57]. The support and confidence values are

calculated to reveal frequently occurring items with Apriori algorithm [58]. To reduce the frequent items, Apriori algorithm includes knowledge on associations between items. In this way, it helps in finding only the related association rules [59].

Table 6 only includes sample subscribers' preferences and indicates that IPTV subscribers prefer comedy genre types, while more than %96 of all other rentals include action, adventure, drama and sci-fi. Additionally, 99% of the customers renting the mentioned genres also rented comedy.

TABLE 6. The relation between comedy and other popular genres for sample subscribers.

Number of rules	lhs	rhs	Support	Confidence	Lift
[1]	{Action=1,Adventure=1}	{Comedy=1}	0.99	0.99	0.99
[2]	{Adventure=1,Drama=1}	{Comedy=1}	0.97	0.99	0.99
[3]	{Action=1,Drama=1}	{Comedy=1}	0.97	0.99	0.99
[4]	{Adventure=1,Sci.Fi=1}	{Comedy=1}	0.96	0.99	0.99
[5]	{Action=1,Sci.Fi=1}	{Comedy=1}	0.96	0.99	0.99

In Figures 3 to 6, green circles illustrate the popular content types rented by the given customer groups; accordingly, the size of these circles indicates the rental volume for each content type, whereas the arrows represent the relationship between the lhs and rhs item set. Likewise, the size of the purple/red circles indicates the lift value; larger circles imply stronger lift.

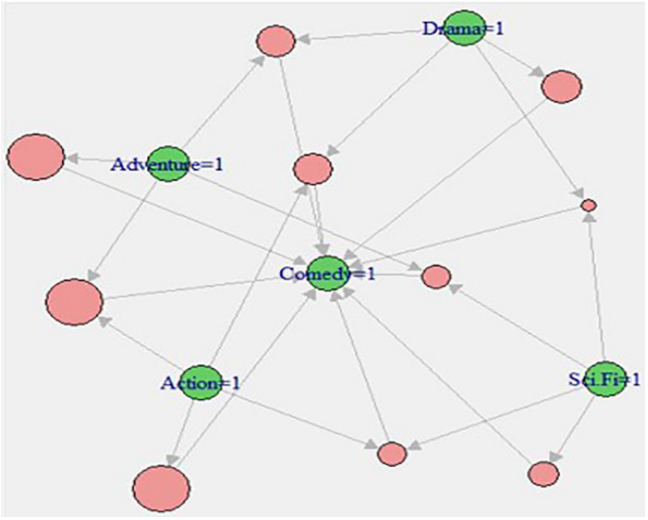

FIGURE 3. The relations among popular content types for "high consuming-valuable subscribers".

According to the analyses, for the "high consuming-valuable subscribers" (cluster 1), the most rented content is found to be comedy, adventure, action, drama and sci-fi in order of preference, thereby suggesting that customers who rent any of these popular genres prefer comedy content, as determined by the Apriori algorithm. Figure 3 represents the result of the relationship between the types of content rented by "high consuming subscribers". In line with what was stated previously, this figure reveals that the lift value is high between the lhs item set (adventure, action, drama and sci-fi) and the rhs item (comedy), thereby implying comedy as a preferred content by those renting adventure, action, drama and sci-fi in this group. To this end, the IPTV service providers could offer the content types relevant to these subscribers' preferences and apply electronic means of notification - emails and other forms of social media - to increase satisfaction in the long term.

As to the "less consuming subscribers" (cluster 2), it can be seen that those renting one of the content types of action, comedy, drama, and animation also chose the adventure genre. As shown in Figure 4, the lift value between lhs (animation and comedy) and rhs (adventure) is higher. This suggests that adventure is a preferred content for those renting animation and comedy in this customer group and, as such, IPTV service providers could add such genres and categories, increase the number of items on offer, and add details to the metadata information to achieve more customer satisfaction.

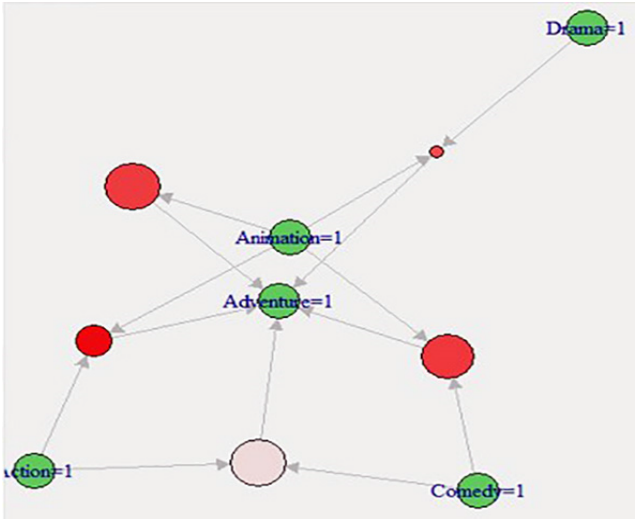

FIGURE 4. The relations among popular content types for "less consuming subscribers".

Subsequently, cluster 3, "less-consuming-loyal subscribers", is found to rent adventure, action, drama and sci-fi, as well as comedy. As shown in Figure 5, the lift values between lhs (action and adventure) and rhs (comedy) are higher – an indication that adventure is a preferred content for those renting animation and comedy. Therefore, IPTV service providers may again employ different notification tools to inform the users of, for example, reduced fees and enriched contents to encourage further rentals.

Lastly, cluster 4 pertaining to "disloyal subscribers" reveals that users who rent action, adventure, drama and sci-fi also opt for comedy – in this way, resembling cluster 1 (high consuming-valuable subscribers) in terms of choosing action, adventure, drama and sci-fi. As shown in Figure-6, the lift values between lhs (adventure, drama and Sci-fi) and rhs (comedy) are higher.; this translates to comedy as

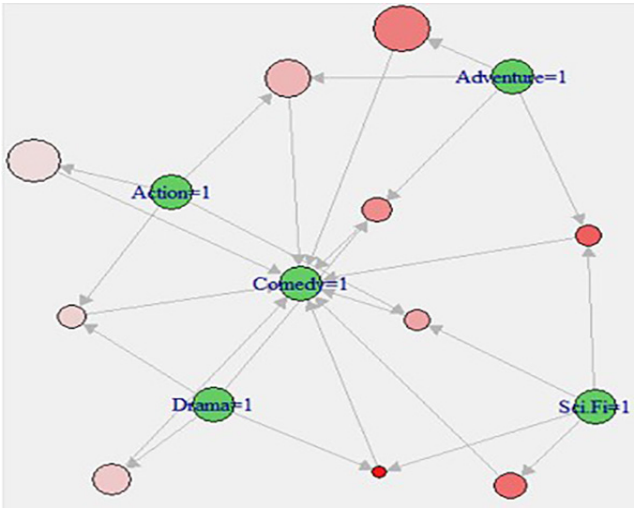

FIGURE 5. The relations among popular content types for "less consuming-loyal subscribers".

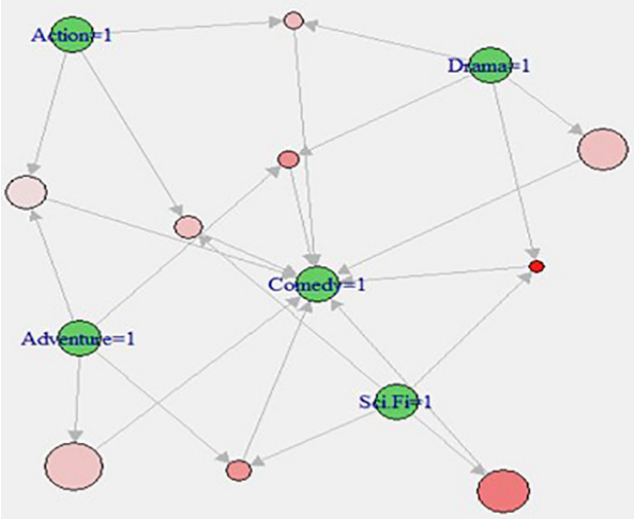

FIGURE 6. The relations among popular content types for "disloyal subscribers".

a preferred content for those renting adventure, drama, and sci-fi in this category. Given the sizeable majority of these subscribers, service providers should carry out more in-depth surveys and analyses to uncover the reasons behind the users' behavior, based on which future marketing decisions and promotional initiatives could be planned.

V. CONCLUSION

With the rapid changes and development of the Internet technology, a surge is witnessed in the number of subscribers using IPTV. Evidently, due to the enormous diversity of such services, the problems related to identifying customers' preferences and expectations are likely to multiply just as well. For this reason, companies should make additional effort to provide appropriate contents to their subscribers by gaining better insights into their behaviors and adopting proper marketing strategies. To achieve these purposes, it is necessary to identify customer values, which will pave the way for improved decision-making and, ultimately, increased profitability for companies in the sector.

With these goals in mind, the proposed combined approach in this study includes clustering and association rule mining for customer profiling in the VoD sector. The LRFMP model employed here allows for the extraction of subscribers' values with the k-means algorithm to carry out the segmentation process. After determining the most suitable cluster number in the study based on real customer data of IPTV service providers, subscribers are grouped into four categories, namely ''high consuming-valuable subscribers'', ''less consuming subscribers'', ''less consuming-loyal subscribers'', and ''disloyal subscribers''. In addition, association rule mining is used to account for the rented content type information. Using the Apriori algorithm, these genres preferred by customers in different clusters are identified.

As for the significance of the present study, it is believed to contribute to the available literature in terms of both theory and practice. This study theoretically contributes to current literature proposing a novel combined data mining approach to the VoD sector to determine different customer groups based on their content preferences, and to analyze these preferences in terms of the prevailing relations among them. Further, this study serves a valuable reference for researchers, who are willing to study field. In this manner, the current research provides insights on how to develop a combined approach using clustering and association rule mining techniques for profiling subscribers in VoD service providers.

From a practical standpoint, the findings of this study contribute in implementing CRM and marketing strategies of IPTV service providers. With these findings, IPTV service providers could offer subscribers appropriate promotions and advertising campaigns for each identified customer profile. Apart from this, more relevant content types and categories could be recommended to users according to their preferences and profiles. Marketing campaigns, in line with customer-base expansion policies, are likely to increase revenue through initiatives such as special-day discounts and promotion of new content via different means of electronic notification as the technology allows today. In short, enhanced and improved strategizing by firms can help them in reaching out to more prospect users while maintaining the existing subscribers in a sustained manner. Lastly, all such developments are likely to change the way IPTV service providers operate in the future as technology enters a new era, a long which customer expectations and profiles may as well evolve.

While the results and findings may be helpful for the further studies, there are a number of limitations in this study. First, the present study makes use of the subscribers' VoD transaction records via only STB devices. In this manner, one direction in the future work is to utilize and investigate the transactional data obtained from other devices such as computers, mobile phones, or the like in the proposed

approach. In future studies, the approach can also be further enhanced by other factors related to customer demographics such as age, gender, profession, income level, region, etc. One more point to consider is that the data were provided by a single case company and limited to Turkish customers. Future research could replicate our study using data from numerous VoD service providers in different geographical areas. All aforementioned possible future directions are important to ensure a more generalized and applicable conclusion. Lastly, the proposed approach employs specific clustering and association rule mining algorithms, and in further studies, alternative clustering and association rule mining algorithms can be used to compare results with this study. Such future studies will be useful to validate the findings drawn from this study.

REFERENCES

[1] T. Arul and A. Shoufan, "Subscription-free pay-TV over IPTV," *J. Syst. Archit.*, vol. 64, pp. 37–49, Mar. 2016.
[2] N. M. Dawi, A. Jusoh, K. M. Nor, and M. I. Qureshi, "Service quality dimensions in pay TV industry: A preliminary study," *Int. Rev. Manage. Marketing*, vol. 6, no. 4S, pp. 239–249, 2016.
[3] I. Kalbandi, M. V. Pawar, B. S. Nikhilkumar, and R. Bachate, "IPTV software process and workflow," *Procedia Comput. Sci.*, vol. 50, pp. 128–134, Jan. 2015.
[4] Research and Markets. (2020). *Video on Demand Market by Solution, Monetization Model, Industry Vertical, and Region—Global Forecast to 2024*. Accessed: Feb. 15, 2020. [Online]. Available: https://www.researchandmarkets.com/reports/4912113/video-on-demand-VoD-market-by-solution-pay-tv
[5] A. Watson. (2019). *Number of Pay TV Subscribers Worldwide From 2010 to 2024*. Accessed: Feb. 20, 2020. [Online]. Available: https://www.statista.com/statistics/825329/pay-tv-subscribers-worldwide/
[6] İ. Kabasakal, "Customer segmentation based on recency frequency monetary model: A case study in E-retailing," *Bilişim Teknolojileri Dergisi*, vol. 13, no. 1, pp. 47–56, Jan. 2020.
[7] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4176–4184, Apr. 2009.
[8] W.-Y. Chiang, "Establishing high value markets for data-driven customer relationship management systems," *Kybernetes*, vol. 48, no. 3, pp. 650–662, Mar. 2019.
[9] W.-Y. Chiang, "Identifying high-value airlines customers for strategies of online marketing systems," *Kybernetes*, vol. 47, no. 3, pp. 525–538, Mar. 2018.
[10] Y.-S. Chen, C.-H. Cheng, C.-J. Lai, C.-Y. Hsu, and H.-J. Syu, "Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment," *Comput. Biol. Med.*, vol. 42, no. 2, pp. 213–221, Feb. 2012.
[11] H. Abbasimehr and M. Shabani, "A new methodology for customer behavior analysis using time series clustering," *Kybernetes*, to be published, doi: 10.1108/K-09-2018-0506.
[12] A. Dursun and M. Caber, "Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis," *Tourism Manage. Perspect.*, vol. 18, pp. 153–160, Apr. 2016.
[13] M. Wedel, and W. A. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*. Boston, MA, USA: Kluwer, 2000.
[14] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "RFM and CLV: Using iso-value curves for customer base analysis," *J. Marketing Res.*, vol. 42, no. 4, pp. 415–430, Nov. 2005.
[15] E. Bilgic, M. Kantardzic, and O. Cakir, "Retail store segmentation for target marketing," in *Advances in Data Mining: Applications and Theoretical Aspects. ICDM*, vol. 9165. Hamburg, Germany: Springer, 2015.
[16] W.-Y. Chiang, "Applying data mining with a new model on customer relationship management systems: A case of airline industry in taiwan," *Transp. Lett.*, vol. 6, no. 2, pp. 89–97, Apr. 2014.
[17] E. C. Malthouse, "Customer relationship management strategy: More important now than ever before," in *The New Advertising: Branding, Content, and Consumer Relationships in the Data-Driven Social Media Era*. Santa Barbara, CA, USA: Praeger, 2016, pp. 111–134.
[18] W. Reinartz, M. Krafft, and W. D. Hoyer, "The customer relationship management process: Its measurement and impact on performance," *J. Marketing Res.*, vol. 41, no. 3, pp. 293–305, Aug. 2004.
[19] C. Gurău, A. Ranchhod, and R. Hackney, "Customer-centric strategic planning: Integrating CRM in online business systems," *Inf. Technol. Manage.*, vol. 4, pp. 199–214, Apr. 2003.
[20] I. Maryani and D. Riana, "Clustering and profiling of customers using RFM for customer relationship management recommendations," in *Proc. 5th Int. Conf. Cyber IT Service Manage. (CITSM)*, Aug. 2017, pp. 1–6.
[21] M. McDonald, and I. Dunbar, *Market Segmentation: How to Do It, How to Profit From It*. Oxford, U.K.: Butterworth-Heinemann, 2004.
[22] C. Chan, "Intelligent value-based customer segmentation method for campaign management: A case study of automobile retailer," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2754–2762, May 2008.
[23] T. Chen, A. Lu, and S.-M. Hu, "Visual storylines: Semantic visualization of movie sequence," *Comput. Graph.*, vol. 36, no. 4, pp. 241–249, Jun. 2012.
[24] C. L. Bauer, "A direct mail customer purchase model," *J. Direct Marketing*, vol. 2, no. 3, pp. 16–24, 1988.
[25] B. Izadi and A. Sabaghinia, "RFM-based e-markets segmentation using self-organizing maps," *J. Econ. Manage.*, vol. 3, no. 12, pp. 86–96, 2014.
[26] A. M. Hughes, "Boosting response with RFM," *Marketing Tools*, vol. 3, no. 3, pp. 4–8, 1996.
[27] A. Sheikh, T. Ghanbarpour, and D. Gholamiangonabadi, "A preliminary study of fintech industry: A two-stage clustering analysis for customer segmentation in the B2B setting," *J. Bus.-Bus. Marketing*, vol. 26, no. 2, pp. 197–207, Apr. 2019.
[28] S. M. S. Hosseini, A. Maleki, and M. R. Gholamian, "Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty," *Expert Syst. Appl.*, vol. 37, no. 7, pp. 5259–5264, Jul. 2010.
[29] Y.-T. Kao, H.-H. Wu, H.-K. Chen, and E.-C. Chang, "A case study of applying LRFM model and clustering techniques to evaluate customer values," *J. Statist. Manage. Syst.*, vol. 14, no. 2, pp. 267–276, Mar. 2011.
[30] D.-C. Li, W.-L. Dai, and W.-T. Tseng, "A two-stage clustering method to analyze customer characteristics to build discriminative customer management: A case of textile manufacturing business," *Expert Syst. Appl.*, vol. 38, no. 6, pp. 7186–7191, Jun. 2011.
[31] J.-T. Wei, S.-Y. Lin, C.-C. Weng, and H.-H. Wu, "A case study of applying LRFM model in market segmentation of a children's dental clinic," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5529–5533, Apr. 2012.
[32] F. Marisa, S. S. S. Ahmad, Z. I. M. Yusof, F. Fachrudin, and T. M. A. Aziz, "Segmentation model of customer lifetime value in small and medium enterprise (SMEs) using K-means clustering and LRFM model," *Int. J. Integr. Eng.*, vol. 11, no. 3, pp. 169–180, Sep. 2019.
[33] H.-C. Chang and H.-P. Tsai, "Group RFM analysis as a novel framework to discover better customer consumption behavior," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14499–14513, Nov. 2011.
[34] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: A case study," *Marketing Intell. Planning*, vol. 35, no. 4, pp. 544–559, Jun. 2017.
[35] H. Jiawei, and M. Kamber, *Data Mining: Concepts and Techniques*, vol. 5. San Francisco, CA, USA: Morgan Kaufmann, 2001.
[36] T. S. Madhulatha, "An overview on clustering methods," *IOSR J. Eng.*, vol. 2, no. 4, pp. 719–725, Apr. 2012.
[37] Y. M. Cheung, "K-means: A new generalized k-means clustering algorithm," *Pattern Recognit. Lett.*, vol. 24, no. 15, pp. 2883–2893, Nov. 2003.
[38] I. Davidson, "Understanding K-means non-hierarchical clustering," SUNY Albany Tech. Rep. 2, 2002.
[39] P. Michaud, "Clustering techniques," *Future Gener. Comput. Syst.*, vol. 13, nos. 2–3, pp. 135–147, Nov. 1997.
[40] Y. Liu, H. Li, G. Peng, B. Lv, and C. Zhang, "Online purchaser segmentation and promotion strategy selection: Evidence from Chinese E-commerce market," *Ann. Oper. Res.*, vol. 233, no. 1, pp. 263–279, Oct. 2015.
[41] H. Hwang, T. Jung, and E. Suh, "An LTV model and customer segmentation based on customer value: A case study on the wireless telecommunication industry," *Expert Syst. Appl.*, vol. 26, no. 2, pp. 181–188, Feb. 2004.
[42] Y. H. Huang, "Video advertisement mining for predicting revenue using random forest," M.S. thesis, Purdue Univ., West Lafayette, IN, USA, 2015.
[43] V. Babaiyan and S. A. Sarfarazi, "Analyzing customers of South Khorasan telecommunication company with expansion of RFM to LRFM model," *J. AI Data Mining*, vol. 7, no. 2, pp. 331–340, 2019.

[44] Z. Kou, "Association rule mining using chaotic gravitational search algorithm for discovering relations between manufacturing system capabilities and product features," *Concurrent Eng.*, vol. 27, no. 3, pp. 213–232, Sep. 2019.
[45] Y. Djenouri, J. C.-W. Lin, K. Norvag, and H. Ramampiaro, "Highly efficient pattern mining based on transaction decomposition," in *Proc. IEEE 35th Int. Conf. Data Eng. (ICDE)*, Macao, Apr. 2019, pp. 1646–1649.
[46] T. Arjannikov and J. Z. Zhang, "An association-based approach to genre classification in music," in *Proc. ISMIR*, 2014, pp. 95–100.
[47] X. Pan, F. Yin, and J. Chai, "Delayering tagging of television programs and association rule mining," in *Proc. IEEE 17th Int. Conf. Comput. Sci. Eng.*, Dec. 2014, pp. 192–197.
[48] S. Hussain, R. Atallah, A. Kamsin, and J. Hazarika, "Classification, clustering and association rule mining in educational datasets using data mining tools: A case study," in *Cybernetics and Algorithms in Intelligent Systems. CSOC*, vol. 765. Vsetin, Czech Republic: Springer, 2018, pp. 196–211.
[49] S. Goh, A. Ron and Z. Shen, *Mathematics and Computation in Imaging Science and Information Processing*. Singapore: World Scientific, 2007.
[50] U. Gürsoy, Ö. A. Kasapoğlu, and K. Atalay, "Association rules analysis with R programming: Analyzing customer shopping data with Apriori and eclat algorithms," *Alphanumer. J.*, vol. 7, no. 2, pp. 357–368, Dec. 2019.
[51] D. J. Ketchen, Jr., and C. L. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *Strategic Manage. J.*, vol. 17, no. 6, pp. 441–458, Jun. 1996.
[52] G. W. Milligan and M. C. Cooper, "A study of standardization of variables in cluster analysis," *J. Classification*, vol. 5, no. 2, pp. 181–204, Sep. 1988.
[53] Y. Liu, K. Yu, X. Wu, Y. Shi, and Y. Tan, "Association rules mining analysis of app usage based on mobile traffic flow data," in *Proc. IEEE 3rd Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2018, pp. 55–60.
[54] S. H. Ha and S. C. Park, "Application of data mining tools to hotel data mart on the intranet for database marketing," *Expert Syst. Appl.*, vol. 15, no. 1, pp. 1–31, Jul. 1998.
[55] C. Marcus, "A practical yet meaningful approach to customer segmentation," *J. Consum. Marketing*, vol. 15, no. 5, pp. 494–504, Oct. 1998.
[56] H. H. Chang and S. F. Tsay, "Integrating of SOM and K-mean in data mining clustering: An empirical study of CRM and profitability evaluation," *J. Inf. Manage.*, vol. 11, no. 4, pp. 161–203, 2004.
[57] Y. Djenouri, D. Djenouri, J. C.-W. Lin, and A. Belhadi, "Frequent itemset mining in big data with effective single scan algorithms," *IEEE Access*, vol. 6, pp. 68013–68026, 2018.
[58] Y. Djenouri and M. Comuzzi, "Combining apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem," *Inf. Sci.*, vol. 420, pp. 1–15, Dec. 2017.
[59] Y. Djenouri, A. Belhadi, and P. Fournier-Viger, "Extracting useful knowledge from event logs: A frequent itemset mining approach," *Knowl.-Based Syst.*, vol. 139, pp. 132–148, Jan. 2018.

SINEM GUNEY is currently pursuing the Ph.D. degree with the Department of Software Engineering, Atilim University, Ankara, Turkey. Also, she is working as a Media Applications Specialist in Turkey's one of largest telecom service providers. Her research interests are software engineering, technological advancements, data mining, and customer segmentation.

SERHAT PEKER received the Ph.D. degree in information systems from Middle East Technical University, Turkey. He is currently an Assistant Professor with the Department of Management Information Systems, İzmir Bakirçay University, Turkey. His research interests are consumer/user behavior modeling and human–computer interaction, data mining, and machine learning.

CIGDEM TURHAN received the Ph.D. degree in computer engineering from the Middle East Technical University, Ankara, Turkey. She is currently working as an Assistant Professor with the Department of Software Engineering, Atilim University, Ankara. She is the author of a number of text books in the area of programming. Her research interests include natural language processing, machine translation, semantic web technologies, and engineering education.